Automatic evaluation of disclosure risks of text anonymization methods

Benet Manzanares-Salor[1[0000-0001-7254-2560]], David Sánchez[1[0000-0001-7275-7887]] and Pierre Lison[2[0000-0002-7649-0689]]

[1] Universitat Rovira i Virgili, Department of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, Tarragona, Spain
[2] Norwegian Computing Center, Oslo, Norway
{benet.manzanares,david.sanchez}@urv.cat, plison@nr.no

Abstract. The standard approach to evaluate text anonymization methods consists of comparing their outcomes with the anonymization performed by human experts. The degree of privacy protection attained is then measured with the IR-based recall metric, which expresses the proportion of re-identifying terms that were correctly detected by the anonymization method. However, the use of recall to estimate the degree of privacy protection suffers from several limitations. The first is that it assigns a uniform weight to each re-identifying term, thereby ignoring the fact that some missed re-identifying terms may have a larger influence on the disclosure risk than others. Furthermore, IR-based metrics assume the existence of a single gold standard annotation. This assumption does not hold for text anonymization, where several maskings (each one encompassing a different combination of terms) could be equally valid to prevent disclosure. Finally, those metrics rely on manually anonymized datasets, which are inherently subjective and may be prone to various errors, omissions and inconsistencies. To tackle these issues, we propose an automatic re-identification attack for (anonymized) texts that provides a realistic assessment of disclosure risks. Our method follows a similar premise as the well-known record linkage methods employed to evaluate anonymized structured data, and leverages state-of-the-art deep learning language models to exploit the background knowledge available to potential attackers. We also report empirical evaluations of several well-known methods and tools for text anonymization. Results show significant re-identification risks for all methods, including also manual anonymization efforts.

Keywords: text anonymization, re-identification risk, language models, BERT.

1 Introduction

The availability of textual data is crucial for many research tasks and business analytics. However, due to its human origin, textual data often includes personal private information. In such case, appropriate measures should be undertaken prior distributing the data to third parties or releasing them to the public in order to comply with the General

Data Protection Regulation (GDPR) [1]. These measures involve either obtaining explicit consent of the individuals the data refer to (which may be infeasible in many cases), or applying an anonymization process by which the data can no longer be attributed to specific individuals. The latter renders data no longer personal and, therefore, outside the scope of the GDPR.

Data anonymization has been widely employed to protect structured databases, in which the individuals' data consist of records of attributes. In this context, a variety of well-stablished anonymization methods and privacy models have been proposed, such as *k*-anonymity and its extensions [2-4], or ε-differential privacy [5]. However, plain (unstructured) text anonymization is significantly more challenging [6, 7]. The challenges derive from the fact that the re-identifying personal attributes mentioned in the text are unbounded and, quite often, not clearly linked to the individual they refer to.
Most approaches to text anonymization rely on natural language processing (NLP) techniques –named entity recognition (NER)– [8-20] to detect and mask words of potentially sensitive categories, such as names or addresses. Since these methods limit masking to (a typically reduced set of) pre-established categories, they usually offer weak protection against re-identification, the latter being caused by a large variety of entity types. Alternately, methods proposed in the area of privacy preserving data publishing (PPDP) [21-27] consider *any* information that jeopardizes individual's anonymity. However, the damage they cause to the data and several scalability issues make them unpractical in many scenarios [6].

Moreover, because most text anonymization methods do not offer formal privacy guarantees, the degree of protection they offer should be empirically evaluated, as done in the statistical disclosure control (SDC) literature [28]. The standard way to evaluate text anonymization methods consists of comparing their outcomes with manually anonymized versions of the documents to be protected [8, 10-15, 18, 20, 21]. The performance of anonymization methods is then measured through IR-based metrics, specifically precision and recall. Whereas precision accounts for unnecessarily masked terms (which would negatively affect the utility and readability of the anonymized outcomes), recall, which accounts for the amount of undetected re-identifying terms, is roughly equaled as the inverse of disclosure risk. However, recall is severely limited because i) not all (missed) re-identifying terms contribute equally to disclosure, ii) several maskings (each one encompassing a different combination of terms) could be equally valid to prevent disclosure, and iii) it relies on manual anonymization, which may be prone to errors and omissions [6, 29].

In contrast, in the SDC field, the disclosure risk of anonymized databases is empirically measured by subjecting the anonymized data to re-identification attacks, more specifically, *record linkage attacks* [30-33]. Record linkage matches records in the protected database and a background database containing publicly available identified information of the protected individuals. Because successful matchings between both databases results in re-identification, the percentage of correct record linkages provides a realistic an objective measure of the disclosure risk, and an accurate simulation of what an external attacker may learn from the anonymized outcomes.

Because assessing disclosure risks by measuring the performance of automatic re-identification attacks is more convenient and realistic than relying on (limited and human-dependent) IR-based metrics, in this paper we propose a re-identification attack for text anonymization methods grounded on the same formal principles as the record linkage attack employed in structured databases. On that basis, we also provide an intuitive disclosure risk metric based on the re-identification accuracy, which overcomes the limitations of the commonly employed recall-based risk assessment.

To maximize re-identifiability, our attack leverages state-of-the-art machine learning techniques for NLP [34]. These techniques have proved to obtain human or above-human level in several language-related tasks, thereby making our method a realistic representation of an ideal human attacker. We also show the application of our attack to evaluate the level of protection offered by a variety of widely used and state-of-the-art text anonymization methods and tools, in addition to a sample of human-based anonymization employed in a previous work as evaluation ground truth [27].

The remainder of this paper is organized as follows. Section 2 discusses related works on privacy evaluation. Section 3 presents our attack and metric for assessing the re-identification risk of anonymized texts. Section 4 reports and discusses the empirical evaluation of a variety of automated and manual anonymization approaches. The final section gathers the conclusions and depicts lines of some future research.

2 Related work

In the context of document anonymization, recall is used as standard to evaluate the level of privacy protection attained by automatic anonymization methods [8, 10-15, 18, 20, 21]. Recall is an IR-based completeness metric, which is defined as the fraction of relevant instances that were properly identified by the method to be evaluated:

$$Recall = \frac{\#TruePositives}{\#TruePositives+\#FalseNegatives} \qquad (1)$$

where *#TruePositives* is the number of relevant instances identified and *#FalseNegatives* represents the missed ones. In text anonymization, the relevant instances correspond to words or n-grams that should be masked. These are identified via manual annotation, which is considered the ground truth.

Because IR-based metrics (precision and recall) are the standard way to evaluate many NLP tasks (and NER in particular), and NER techniques are the most common way to tackle text anonymization, perhaps by inertia, the vast majority of methods employ recall to assess the level of attained privacy protection. Nevertheless, this suffers from a variety of issues [29, 35]. First, recall does not measure the actual residual disclosure risk of anonymized documents, but just compares the outputs with manual annotations. Manual anonymization is by definition, subjective and non-unique, and may be prone to errors, bias and omissions [6, 29]. On top of that, manual annotation is costly and time consuming, and usually involves several human experts, whose annotations should be integrated through a non-trivial process. Another limitation of recall-based evaluation is that it assumes that all identified/missed entities contribute equally

to mitigate/increase the risk, which is certainly inaccurate [29]. Obviously, failing to mask identifying information (such as a proper name) is much more disclosive on the individual to be protected than just missing her job or her whereabouts.

On the other hand, in the area of SDC, the level of privacy protection attained by anonymization methods on a structured database is measured according to the success of a re-identification attack (*record linkage* [33]) that a hypothetical attacker could perform on the anonymized outcomes. Record linkage tries re-identify anonymized records by linking the masked quasi-identifiers present in those records with those available on publicly available identified sources. Then, the re-identification risk is measured as the percentage of correct linkages:

$$Re\text{-}identification\ risk \approx Linkage\ accuracy = \frac{\#CorrectLinkedRecords}{\#Records} \quad (2)$$

Compared to recall, the record linkage accuracy offers an automatic and objective means to evaluate privacy that does not rely on manual annotations.

3 A re-identification attack for evaluating anonymized text

In this section, we present a re-identification attack for (anonymized) text based on state-of-the-art NLP machine learning techniques. Our attack aims to provide a practical, realistic and objective mean to evaluate the privacy protection offered by anonymization methods for textual data.

In broad terms, the attack aims to re-identify the individuals referred in a set of anonymized documents by leveraging a classifier trained on a collection of identified and publicly available documents encompassing a population of subjects in which the individuals referred in the anonymized documents are contained. For example, one may use publicly available social media publications from a city's inhabitants to re-identify anonymized medical reports from that city's hospital. By construction, the publicly available data should be a superset of the anonymized set. The protected documents would contain confidential attributes (e.g., diagnoses) and masked quasi-identifiers (e.g., age intervals) from unidentified individuals, whereas the publicly available documents would contain identifiers (e.g., a complete name) and clear quasi-identifiers (e.g., a specific age) from known individuals. Consequently, unequivocal matchings of the (quasi-)identifiers of both types of documents (due to a weak anonymization), would allow re-identifying the protected documents and, therefore, disclose the confidential attributes of the corresponding individuals.

Our method can be seen as an adaptation of the standard record linkage attack from structured databases to textual data, where documents correspond to records, words (or n-grams) roughly correspond to attribute values and the classifier provides the criterion to find the best match/linkage between the anonymized and public documents.

The attack is designed with the aim of recreating as realistically as possible what a real attacker would do to re-identify the protected individuals. This also accounts for the amount of resources (computation and background data) that a real attacker may reasonably devote and have available to execute the attack. This is in line with the

GDPR (Recital 26), which specifies that, to assess the risk of re-identification, one should take into account the reasonable means that can be employed to perform such re-identification. This makes our attack and the derived risk metric more realistic.

Formally, let A_D be the set of anonymized (non-identified) documents and B_D the set of identified publicly available documents (i.e., background documents). Each document describes or refers to a specific individual, thereby defining the sets of individuals A_I and B_I, and the mapping bijective functions F_A: $A_D \rightarrow A_I$ and F_B: $B_D \rightarrow B_I$. Assuming $A_I \subseteq B_I$ (as in the original record linkage attack), F_C: $A_D \rightarrow B_I$ is the re-identification function that matches protected documents with the corresponding known individuals. On this basis, from the point of view of an attacker, A_D, B_D, B_I and F_B are known, and A_I, F_A and F_C are unknown. Therefore, the purpose of the attack is obtaining F_C' (an approximation of F_C) by exploiting the similarities between A_D and B_D sets.

In Algorithm 1 we formalize our proposal, which returns the number of correct re-identifications achieved by the attack on an input collection of anonymized documents. First, a machine learning classifier is built and trained to predict F_C (line 1, more details in Section 3.1). Using the formal notation above, the classifier would implement F_C' by learning which individuals from B_I correspond to the documents in B_D according to the knowledge available to the attacker. Subsequently, the same classifier is evaluated with the set of anonymized documents A_D (line 4). A correct re-identification would happen if the prediction (i.e., F_C') matches F_C (lines 5-6). Finally, the number of re-identifications are returned (line 9).

Algorithm 1. Re-identification risk assessment for anonymized text documents

```
Input: A_D // set of anonymized documents
       B_D  // set of background documents
       B_I  // set of individuals from background documents
       F_B  // mapping function from B_D to B_I
       F_C  // groundtruth mapping function from A_D to B_I
Output: numReIds // number of correct re-identifications

1  classf = build_classifier(B_D, B_I, F_B, A_D);
2  numReIds = 0; // Number of correct re-identifications
3  for each d in A_D do // Evaluation loop for all documents
4     pred_ind = classf.predict(d); // Predicted B_I individual for d
5     if (pred_ind == F_C(d)) then // If correct re-identification
6        numReIds++;
7     end if
8  end for
9  return numReIds;
```

Similarly to the record linkage method (Eq. 2), we assess the re-identification risk of A_D according to the accuracy of the re-identification attack:

$$Re\text{-}identification\ risk \approx Re\text{-}identification\ accuracy = \frac{numReIds}{|A_I|} \quad (3)$$

3.1 Building the classifier

We next detail the internals of the *build_classifier* method (line 1 of Algorithm 1). Its goal is to reproduce as faithfully as possible the techniques that a potential attacker may employ to conduct the re-identification attack. This includes considering state-of-the-art NLP classification models and taking advantage of the data available to the attacker. To select the model, we consider state-of-the-art *word embedding* and *transformer*-based models, which have recently revolutionized the area of NLP. Word embeddings [36] map words (tokens) to real-valued vector representations that capture their meaning, so that words closer in the vector space are expected to be semantically related. The initial approaches to word embeddings produced a fixed vector for each token. Nevertheless, in many cases, words' meaning is affected by the context (especially for polysemic words) and, therefore, they cannot be properly defined through unique embeddings. This led to the creation of contextual word embeddings [37], where the embedding depends on the context of the word instance. Since our classifier requires non-ambiguous words representations, which allow to determine if a word is related with a particular individual, using contextual word embeddings is the best strategy.

Word embedding models require from large training corpora in order to build general and robust word representations. This has led to the popularization of *pre-trained* models [34, 38], which are trained once with an enormous corpus and then are used in multiple NLP tasks. Even though the results obtained from these pre-trained models are good enough for a variety of problems, better performance can be achieved through *fine-tuning*, a procedure in which word embeddings are further trained with the task's specific corpus. We expect the attacker to follow this paradigm, which provides high quality results while significantly reducing cost of training models from scratch.

Another technology that took a step forward in NLP is the *transformer* architecture [39]. The strengths of this approach are the capability of handling long-range dependencies with ease and a reduced processing time based on parallelism. One of the most popular and well-established transformer-based model for NLP is BERT (Bidirectional Encoder Representations from Transformers) [34], which is pre-trained with a huge corpora (Wikipedia and the BookCorpus), and is capable of learning high quality contextual word embeddings. After simple modifications and fine-tuning, BERT is capable of obtaining human-level or even better performance in multiple language-related tasks, including document classification. On this basis, we consider BERT (or its variations) a well-suited model for our attack, since it can obtain outstanding results with neither a huge cost nor unfeasible knowledge assumptions from the attacker.

In addition to build her own classifier, we also expect the attacker to define a development set to have an intuition of the classifier's performance. In this way, it would be also possible to tune the classifier's hyperparameters to maximize the re-identification accuracy. This configures training as a best model search, in which multiple hyperparameters are evaluated according to the accuracy obtained on the development set.

Going back to our algorithm, the classifier returned by the *build_classifier* method is such that, after the further pre-training and fine-tuning steps, obtains the best accuracy on the development set. To this end, multiple trainings with different hyperparam-

eters are performed, searching the best combination. A fixed number of epochs is defined for further pre-training, and fine-tuning is run until a pre-defined maximum number of epochs is achieved or development accuracy does not improve (early stopping).

Regarding the data that can be employed to build the classifier and the development set, recall that the attacker knowledge is limited to B_D, B_I, F_B and A_D. On the one hand, documents in B_D provide knowledge of the individuals' specific vocabulary, which improves understanding of domain-specific words. Additionally, B_D can be labeled on B_I by using F_B, thereby providing useful information about the relationship between the publicly available background data and the individuals' identity. This can lead to the detection of (quasi-)identifying attributes (e.g., the person's name or her demographic attributes), which are the base of the re-identification attack. On the other hand, unlabeled documents in A_D convey knowledge on the anonymized vocabulary. This includes information such as the co-occurrence of words left in clear with those subjected to masking, which may allow inferring the latter from the former.

On this basis, a straightforward approach would be to use all documents in B_D and A_D for further pre-training, and documents in B_D labeled on B_I for fine-tuning. This produces a model with domain-specific knowledge capable of mapping documents to B_I, as it is required for the attack. Nonetheless, it is important to note that the goal of the model is to correctly classify documents in A_D, which come from a different data distribution than the documents in B_D. Concretely, B_D are clear texts (such as identified posts in social media) whereas A_D are anonymized texts (such as non-identified medical reports with some words masked via suppression or generalization). Because machine learning algorithms are sensitive to differences between training and test data distributions, this could hamper the accuracy. For example, during the fine-tuning step, the classifier may learn to focus on identifying words or structures that are not present in the anonymized documents, which would be useless for the attack. To tackle this problem, we propose creating an anonymized version of B_D called B_D' by using any off-the-shelf text anonymization method available to the attacker. Ideally the same method used for A_D should be employed but, because such method would be usually unknown, a standard NER-based method (being NER the most common approach for practical text anonymization), can be used instead. As a result, documents in B_D' would provide an approximation of how data are anonymized, by employing documents more similar to those in A_D. This offers useful information on how known documents (B_D) are anonymized, thereby facilitating disclosure of masked words based on their context. In addition, B_D' can be labeled on B_I (since B_D' $\rightarrow$ B_D is known), therefore facilitating the discovery of the identities underlying the masked documents; for instance, by discovering identifying words neglected by the anonymization method (e.g., a particular street name) that are also present in documents from A_D. Taking this into consideration, we propose using B_D, B_D' and A_D documents for further pre-training and the union of B_D and B_D' labeled on B_I for fine-tuning, thereby obtaining a classifier model better adapted to the content of the anonymized documents.

For the development set, we propose to extract a random subset of configurable size from the documents in B_D, which we call C_D, and transform it to match, as much as possible, the data distribution of A_D. An intuitive approach would be to anonymize C_D; however, this would result into identical documents to those in B_D', which are already

present in training data. Thereupon, a previous step is required, aiming to differentiate C_D texts from the B_D ones and, if possible, to assimilate them to those in A_D prior anonymization. To this end, we propose to perform a summarization-like process on documents from C_D, obtaining $\hat{C}_D$. On this basis, abstractive or hybrid summarization methods are preferred rather than extractive ones [40], so that they produce summarizations that do not include sentences present in documents from B_D. After that, the summarized documents in $\hat{C}_D$ are anonymized (obtaining $\hat{C}_D$') in the same way as done for B_D'. Finally, the documents in $\hat{C}_D$' are used as the development set of the attack.

4 Empirical experiments

This section reports empirical results on the application of our re-identification attack to a variety of text anonymization methods, both NLP-oriented and PPDP-grounded. We also test the risk resulting from a manual anonymization effort.

As introduced above, NLP methods [8-20] tackle anonymization as a NER task, in which allegedly private information categories (names, locations, dates, etc.) are detected and masked. Detection is based on rules and models trained to identify the specific categories, and masking consists of replacing the detected entities by their corresponding categories. We considered the following systems and tools that have been employed for NER-based text anonymization [6]:

- *Stanford NER* [41]: provides three pre-trained NER models: *NER3*, which detects ORGANIZATION, LOCATION and PERSON types; *NER4*, which adds the MISC (miscellaneous) type; and *NER7*, which detects ORGANIZATION, DATE, MONEY, PERSON, PERCENT and TIME types.
- *Microsoft Presidio*[1]: a NER-based tool specifically oriented towards anonymization. Among the variety of types supported by Presidio, we enabled those corresponding to quasi-identifying information: NRP -person's nationality, religious or political group-, LOCATION, PERSON and DATE_TIME types.
- *spaCy NER*[2]: we used the *en_core_web_lg*, model, which is capable of detecting named entities of CARDINAL, DATE, EVENT, FAC (e.g., buildings, airports, etc.), GPE (e.g., countries, cities, etc.), LANGUAGE, LAW (named documents made into laws), LOC (non-GPE locations such as mountain ranges), MONEY, NORP (nationalities or religious political group), ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME and WORK_OF_ART types.

Regarding PPDP text anonymization methods, most of them are on the theoretical side [23, 25, 26], suffer from severe scalability issues [21, 42, 43] or seriously damage data utility [22, 24], making them hardly applicable. The only practical method we found is [27], which is based on word embedding models. Due to the lack of a name, this method will be referred to as Word2Vec, this being the backbone neural model employed by this work.

[1] https://github.com/microsoft/presidio
[2] https://spacy.io/api/entityrecognizer

In addition to automatic methods, we also considered the manual anonymization conducted by the authors of [27], which allows us to assess the robustness of manual effort against our re-identification attack. Finally, we also report re-identification results on the unprotected versions of the documents in A_D. This constitutes the baseline risk that anonymization methods should (significantly) reduce.

As evaluation data, we employed the corpus described in [27], which consists of 19,000 Wikipedia articles under the "20th century actors" category. To simulate the scenario described in Section 3, we considered the article abstracts as the private documents to be anonymized, whereas the article bodies (whose content overlap with the abstracts, even though presented in a different, more detailed way) were assumed to be the identified publicly available information. From this corpus, 50 article abstracts corresponding to popular, contemporary and English speaking actors were extracted in [27] as the set to be subjected to both automatic and manual anonymization. In terms of our attack, the 50 actors in the extracted set constitute A_I, the 50 abstracts anonymized with a method m define A_D^m, and the article bodies in the corpus constitute B_D (with a population of B_I actors that should encompass A_I).

The amount of background documents B_D used to perform the attack, and their overlap with A_I, have a critical role in the success of the attack. To test this aspect, we defined several attack scenarios by setting increasingly larger B_Ds:

- *50_eval*: a worst case scenario for privacy, in which B_I exactly matches A_I, thereby constituting the easiest re-identification setting. In this case B_D comprises the 50 article bodies of the 50 anonymized abstracts.
- *500_random*: a synthetic scenario consisting of 500 random article bodies taken from the total of 19,000 in the corpus plus those corresponding to the 50 actors in A_I that were not included in the initial random selection. This ensures that $A_I \subseteq B_I$.
- *500_filtered*: a set of 581 article bodies obtained by systematically filtering the initial 19,000 according to several features related to the actors in A_D. In particular, we discarded non-native English speakers, non-actors (e.g., directors), dead individuals, those born before 1950 or after 1995 (latter included) and those whose article included less than 100 links and was present in less than 40 languages (the latter two being related to the 'popularity' of the actor). These criteria aim to maximize the number of individuals in A_I present in B_I, even without knowing A_I, as it would happen in practice. As a result, 40 out of the 50 actors in A_I appeared in B_I. This limits the re-identification accuracy to 80%.
- *2000_filtered*: a set of 1,952 article bodies obtained by using the same criteria as in the prior set but omitting the filter on the number of languages. This results in 41 actors from A_I appearing in B_I, which limits the re-identification accuracy to 82%.

Once B_D is set for a particular scenario, the corresponding B_D', C_D, $\hat{C}_D$ and $\hat{C}_D'$ sets required to define the training and development sets should be created as detailed in Section 3.1. To create B_D', we anonymized the documents in B_D by using spaCy NER. On the other hand, $\hat{C}_D$ comprised a subset of the abstracts corresponding to the bodies in B_D. Being the abstracts summaries of the article bodies, this procedure follows the summarization-based approach proposed in Section 3.1, thus not requiring explicitly building C_D. The size of $\hat{C}_D$ was set to 10% for the *2000_filtered*, *500_filtered* and

500_random scenarios, and 30% for *50_eval*. Finally, the documents in $\hat{C}_D$ were anonymized by following the same method employed for B_D', thus obtaining the $\hat{C}_D$' set that constitutes the development set.

To realistically simulate the implementation of our method by a potential attacker, we considered the resources that such attacker would reasonably devote. On this basis, we employed Google Colaboratory, which offers the most powerful free platform for building and running machine learning models. Resources at Google Colaboratory may vary depending on the actual demand. In our tests, the running environment consisted of an Nvidia Tesla K80 GPU with 16GB of VRAM, an Intel Xeon CPU and 12GB of RAM. Google Colaboratory's free tier limits the maximum duration of a run to 12 hours. Trainings with a longer duration require from saving the current model and manually restoring the process, resulting in a new environment with a potentially different hardware allocation. In order to ensure that all the computation is made on the same hardware (and also to avoid the tedious manual restoring of the test), we didn't consider scenarios with training runtimes longer than 12 hours. This discarded a potential scenario using the whole 19,000 articles as B_D, whose fine-tuning runtime is estimated at about 21 hours for 10 epochs. The other scenarios had runtimes of 31, 99, 297 and 301 minutes, respectively. Note that *500_filtered* took 2.5 times longer to train than *500_random* because the length of the documents in the former was 3 times larger, since the popularity filters applied resulted in longer articles.

Out of the wide variety of pre-trained models based on BERT[3], we have considered those that stand out for their accuracy and/or efficiency, and that can be fine-tuned with the limitations of our execution environment (e.g., GPU memory). Under this premise, we selected DistilBERT (*distilbert-base-uncased*), a distilled version of the original BERT which reduces 40% the model's size but keeps a 97% of its performance in multiple tasks; this provides a great trade-off between accuracy and cost.

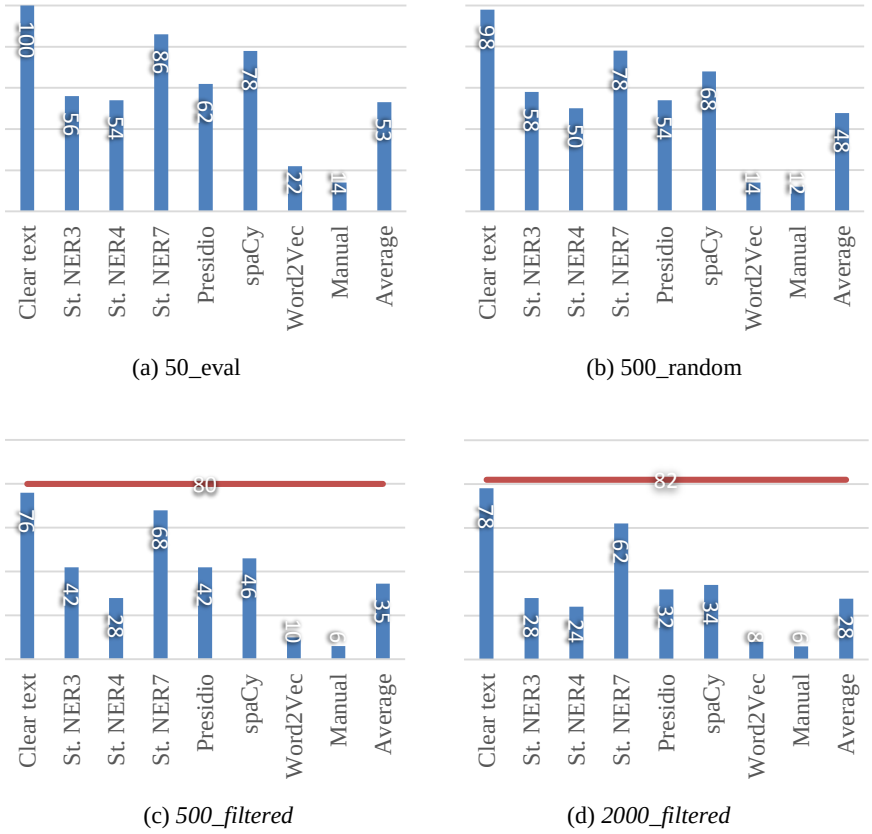
As discussed in Section 3.1, the model training included performing a best model search based on model's hyperparameters. Considering the number of tests to be conducted, their runtime and their similarities, we applied it to the *50_eval* scenario and used the obtained parameters in the remaining scenarios. Specifically, the hyperparameters that provided the best accuracy for the development set were: *learning rate* 5e-5, *batch size* 16, *sliding window length/overlap* 512/128 and *sliding window length/overlap for classification* 100/25. Additionally, the Hugging Face's AdamW optimizer was used with default parameters except for the learning rate (*betas* 0.9 and 0.999, *eps* 1e-8 and *weight decay* 0).

Pre-training was performed during 3 epochs and fine-tuning during a maximum of 20 epochs. Using the accuracy at the development set for early stopping criteria with a patience of 5 epochs, fine-tuning was run for ~20 epochs for the *50_eval*, *500_random* and *500_filtered* scenarios and during ~10 epochs for the *2000_filtered* scenario. Additionally, it is important to note that the pre-training only used B_D and B_D' without performing the optimal fine-tuning using each one of the A_Ds. Doing so would increase the number of tests by a factor of 8 (the number of methods/configurations tested), and we observed no noticeable benefits in the worst-case scenario *50_eval*.

[3] https://huggingface.co/docs/transformers/index

4.1 Results

Fig. 1 depicts the re-identification risk of each combination of background knowledge and anonymization approach.



(a) 50_eval (b) 500_random

(c) *500_filtered* (d) *2000_filtered*

Fig. 1. Re-identification risk percentages of several anonymization approaches with different sets of background documents. In (c) and (d) the maximum possible re-identification accuracy is depicted as a horizontal line.

First, we notice that the re-identification risk of $A_D^{Clear\ text}$ (that is, non-anonymized documents) is close to the maximum, which is, 100% for *50_eval* and *500_random*, and 80% and 82% for *500_filtered* and *2000_filtered*, respectively. This proves the effectiveness of the tuned DistilBERT model as classifier. For the case of anonymized documents, we observe that the attack is capable of re-identifying individuals even from A_D^{Manual}, with accuracies well-above the random guess, which is 2% for *50_eval*, 0.2% for *500_random*, 0.17% for *500_filtered* and 0.05% for *2000_filtered*. This illustrates that manual anonymization efforts are prone to errors and omissions, and are limited when used as evaluation ground truth.

On the other hand, the average re-identification risk illustrate how B_D influences the results. In particular, the *500_random* scenario provides just slightly less re-identification risk than *50_eval*, because the common features of the 50 protected individuals make them easily differentiable within the random set. In contrast, the risk of the filtered $B_D s$ is significantly lower because i) not all the protected individuals are present in B_D and ii) those present are more similar to the other individuals in B_D, thereby being harder to discriminate.

Regarding the different anonymization methods, NER-based techniques show significant deficiencies, reaching re-identification risks greater than 50% for the *50_eval* worst-case scenario and, still, no lower than 20% for *2000_filtered*. On the other hand, the PPDP approach from [27] achieved the best results of any automated method across all $B_D s$, with a re-identification risk just slightly greater than the manual anonymization. That fact that this method does not limit masking to a pre-defined set of categories (as NER-based methods do) certainly contributes to better mimic the human criteria and decrease the disclosure risk.

5 Conclusions and future work

We have proposed an attack-based disclosure risk assessment method for evaluating text anonymization methods. Compared to the standard recall-based privacy evaluation employed in the literature, our method offers an objective, realistic and automatic alternative that does not require costly and time consuming manual annotations. The experimental results we report provide empirical evidences to the criticisms raised in [6, 27] on the limitations of NER-based methods for text anonymization. Our results also suggest that privacy-grounded methods based on state-of-the-art language models (such as the approach in [27]) offer more robust anonymization that better mimics the criteria of human experts. Nevertheless, the reported re-identification accuracies, which are significantly greater than the random guess, suggest that there is still room for improvement, even for manual anonymization.

As future work, we plan to evaluate the influence of the different hypermarameters in the re-identification accuracy and training runtime and, also, test the behavior of other pre-trained models. Furthermore, we plan to compare our re-identification risk assessment to the standard recall metric.

Acknowledgements

Partial support to this work has been received from the Norwegian Research Council (CLEANUP project, grant nr. 308904), the European Commission (projects H2020-871042 “SoBigData++” and H2020-101006879 “MobiDataLab”) and the Government of Catalonia (ICREA Acadèmia Prize to D. Sánchez).

References

1. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data and Repealing Directive 95/46/EC. In: Commission, E. (ed.), (2016)
2. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd International Conference on Data Engineering, pp. 106-115. IEEE, (2007)
3. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1, 3-es (2007)
4. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 557-570 (2002)
5. Dwork, C.: Differential privacy. In: International Colloquium on Automata, Languages, and Programming, pp. 1-12. Springer, (2006)
6. Lison, P., Pilán, I., Sánchez, D., Batet, M., Øvrelid, L.: Anonymisation models for text data: State of the art, challenges and future directions. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4188-4203. (2021)
7. Csányi, G.M., Nagy, D., Vági, R., Vadász, J.P., Orosz, T.: Challenges and Open Problems of Legal Document Anonymization. Symmetry 13, 1490 (2021)
8. Aberdeen, J., Bayer, S., Yeniterzi, R., Wellner, B., Clark, C., Hanauer, D., Malin, B., Hirschman, L.: The MITRE Identification Scrubber Toolkit: design, training, and assessment. International journal of medical informatics 79, 849-859 (2010)
9. Chen, A., Jonnagaddala, J., Nekkantti, C., Liaw, S.-T.: Generation of surrogates for de-identification of electronic health records. MEDINFO 2019: Health and Wellbeing e-Networks for All, pp. 70-73. IOS Press (2019)
10. Dernoncourt, F., Lee, J.Y., Uzuner, O., Szolovits, P.: De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association 24, 596-606 (2017)
11. Johnson, A.E., Bulgarelli, L., Pollard, T.J.: Deidentification of free-text medical records using pre-trained bidirectional transformers. In: Proceedings of the ACM Conference on Health, Inference, and Learning, pp. 214-221. (2020)
12. Liu, Z., Tang, B., Wang, X., Chen, Q.: De-identification of clinical notes via recurrent neural network and conditional random field. Journal of biomedical informatics 75, S34-S42 (2017)
13. Mamede, N., Baptista, J., Dias, F.: Automated anonymization of text documents. In: 2016 IEEE congress on evolutionary computation (CEC), pp. 1287-1294. IEEE, (2016)
14. Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S., Samore, M.H.: Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC medical research methodology 10, 1-16 (2010)
15. Neamatullah, I., Douglass, M.M., Lehman, L.-W.H., Reisner, A., Villarroel, M., Long, W.J., Szolovits, P., Moody, G.B., Mark, R.G., Clifford, G.D.: Automated de-identification of free-text medical records. BMC medical informatics and decision making 8, 1-17 (2008)
16. Reddy, S., Knight, K.: Obfuscating gender in social media writing. In: Proceedings of the First Workshop on NLP and Computational Social Science, pp. 17-26. (2016)
17. Sweeney, L.: Replacing personally-identifying information in medical records, the Scrub system. In: Proceedings of the AMIA annual fall symposium, pp. 333. American Medical Informatics Association, (1996)

18. Szarvas, G., Farkas, R., Busa-Fekete, R.: State-of-the-art anonymization of medical records using an iterative machine learning framework. Journal of the American Medical Informatics Association 14, 574-580 (2007)
19. Xu, Q., Qu, L., Xu, C., Cui, R.: Privacy-aware text rewriting. In: Proceedings of the 12th International Conference on Natural Language Generation, pp. 247-257. (2019)
20. Yang, H., Garibaldi, J.M.: Automatic detection of protected health information from clinic narratives. Journal of biomedical informatics 58, S30-S38 (2015)
21. Sánchez, D., Batet, M.: C-sanitized: A privacy model for document redaction and sanitization. Journal of the Association for Information Science and Technology 67, 148-163 (2016)
22. Mosallanezhad, A., Beigi, G., Liu, H.: Deep reinforcement learning-based text anonymization against private-attribute inference. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2360-2369. (2019)
23. Chakaravarthy, V.T., Gupta, H., Roy, P., Mohania, M.K.: Efficient techniques for document sanitization. In: Proceedings of the 17th ACM conference on Information and knowledge management, pp. 843-852. (2008)
24. Fernandes, N., Dras, M., McIver, A.: Generalised differential privacy for text document processing. In: International Conference on Principles of Security and Trust, pp. 123-148. Springer, Cham, (2019)
25. Cumby, C., Ghani, R.: A machine learning based system for semi-automatically redacting documents. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1628-1635. (2011)
26. Anandan, B., Clifton, C., Jiang, W., Murugesan, M., Pastrana-Camacho, P., Si, L.: t-Plausibility: Generalizing words to desensitize text. Trans. Data Priv. 5, 505-534 (2012)
27. Hassan, F., Sanchez, D., Domingo-Ferrer, J.: Utility-Preserving Privacy Protection of Textual Documents via Word Embeddings. IEEE Transactions on Knowledge and Data Engineering (2021)
28. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., De Wolf, P.-P.: Statistical disclosure control. Wiley New York (2012)
29. Pilán, I., Lison, P., Øvrelid, L., Papadopoulou, A., Sánchez, D., Batet, M.: The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. arXiv preprint arXiv:2202.00443 (2022)
30. Domingo-Ferrer, J., Torra, V.J.S., Computing: Disclosure risk assessment in statistical microdata protection via advanced record linkage. 13, 343-354 (2003)
31. Nin Guerrero, J., Herranz Sotoca, J., Torra i Reventós, V.: On method-specific record linkage for risk assessment. In: Proceedings of the Joint UNECE/Eurostat work session on statistical data confidentiality, pp. 1-12. (2007)
32. Torra, V., Abowd, J.M., Domingo-Ferrer, J.: Using Mahalanobis distance-based record linkage for disclosure risk assessment. In: International Conference on Privacy in Statistical Databases, pp. 233-242. Springer, (2006)
33. Torra, V., Stokes, K.J.I.J.o.U., Fuzziness, Systems, K.-B.: A formalization of record linkage and its application to data protection. 20, 907-919 (2012)
34. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
35. Mozes, M., Kleinberg, B.J.a.p.a.: No Intruder, no Validity: Evaluation Criteria for Privacy-Preserving Text Anonymization. (2021)
36. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of machine learning research 3, 1137-1155 (2003)

37. Liu, Y., Liu, Z., Chua, T.-S., Sun, M.: Topical word embeddings. In: Twenty-ninth AAAI conference on artificial intelligence. (2015)
38. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems, pp. 5998-6008. (2017)
40. El-Kassas, W.S., Salama, C.R., Rafea, A.A., Mohamed, H.K.: Automatic text summarization: A comprehensive survey. Expert Systems with Applications 165, 113679 (2021)
41. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60. (2014)
42. Sánchez, D., Batet, M.: Toward sensitive document release with privacy guarantees. Engineering Applications of Artificial Intelligence 59, 23-34 (2017)
43. Staddon, J., Golle, P., Zimny, B.: Web-Based Inference Detection. In: usenix security symposium. (2007)